07-normalisation-regression.Rmd

Normalising data and visualising trends

Dmytro Fishman

12 April 2016

Authors: **Dmytro Fishman**

Learning Objectives

- Understand the need and idea of data normalisation
- Fit the regression line to the data from global climate trends
- Draw the figure

Load required packages

```
# plotting package
library(ggplot2)
# piping / chaining
library(magrittr)
# modern dataframe manipulations
library(dplyr)
```

```
#>
#> Attaching package: 'dplyr'
#>
#> The following objects are masked from 'package:stats':
#>
#>     filter, lag
#>
#> The following objects are masked from 'package:base':
#>
#>     intersect, setdiff, setequal, union
```

Let's use the same data we had for previous lessons

```
temperature_raw <- read.csv('data/temperature.csv')
```

Data preprocessing from previous lessons in one block

```
temperature_complete <- temperature_raw %>%
                    filter(!is.na(City)) %>%
                    filter(!is.na(Country)) %>%
                    filter(!is.na(AverageTemperatureFahr)) %>%
                    filter(!is.na(AverageTemperatureUncertaintyFahr))
temperature_complete <- select(temperature_complete, -(day))

temperature_complete <- temperature_complete %>%
```

```
  mutate(AverageTemperatureCelsius = (AverageTemperatureFahr-32)*(5/9)) %>%
  mutate(AverageTemperatureUncertaintyCelsius = (AverageTemperatureUncertaintyFahr-32)*(5/9))

temperature_complete$AverageTemperatureFahr <- NULL
temperature_complete$AverageTemperatureUncertaintyFahr <- NULL
head(temperature_complete)
```

Focus on temperature from Ukraine with the use of `filter`

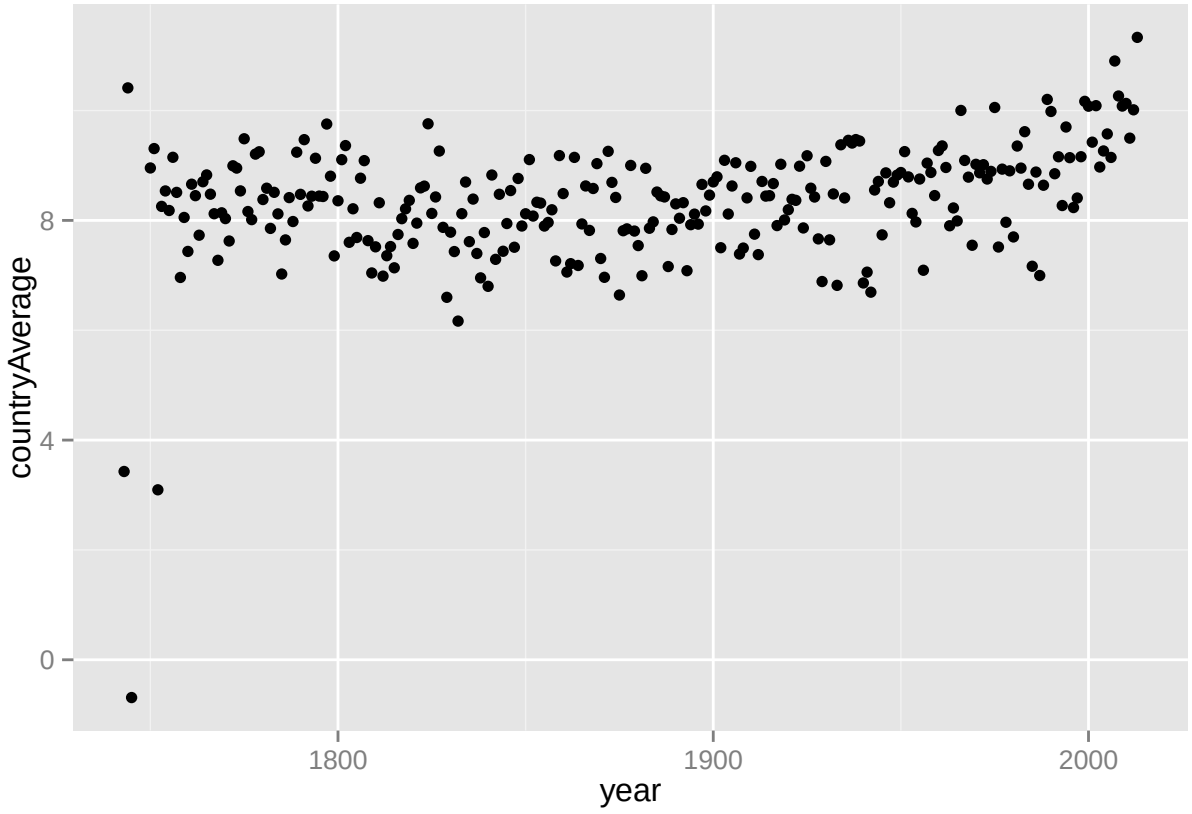
```
temperature_ukraine  <- temperature_complete %>%
                        filter(Country == 'Ukraine')
```

Find mean temperature per each year using `group_by` and `summarise` functions

```
yearly_ukraine_temp <- temperature_ukraine %>%
                        group_by(year) %>%
                        summarise(countryAverage = mean(AverageTemperatureCelsius))
```

Plot this information using `geom_point`

```
ggplot(data = yearly_ukraine_temp, aes(x = year, y = countryAverage)) +
      geom_point()
```



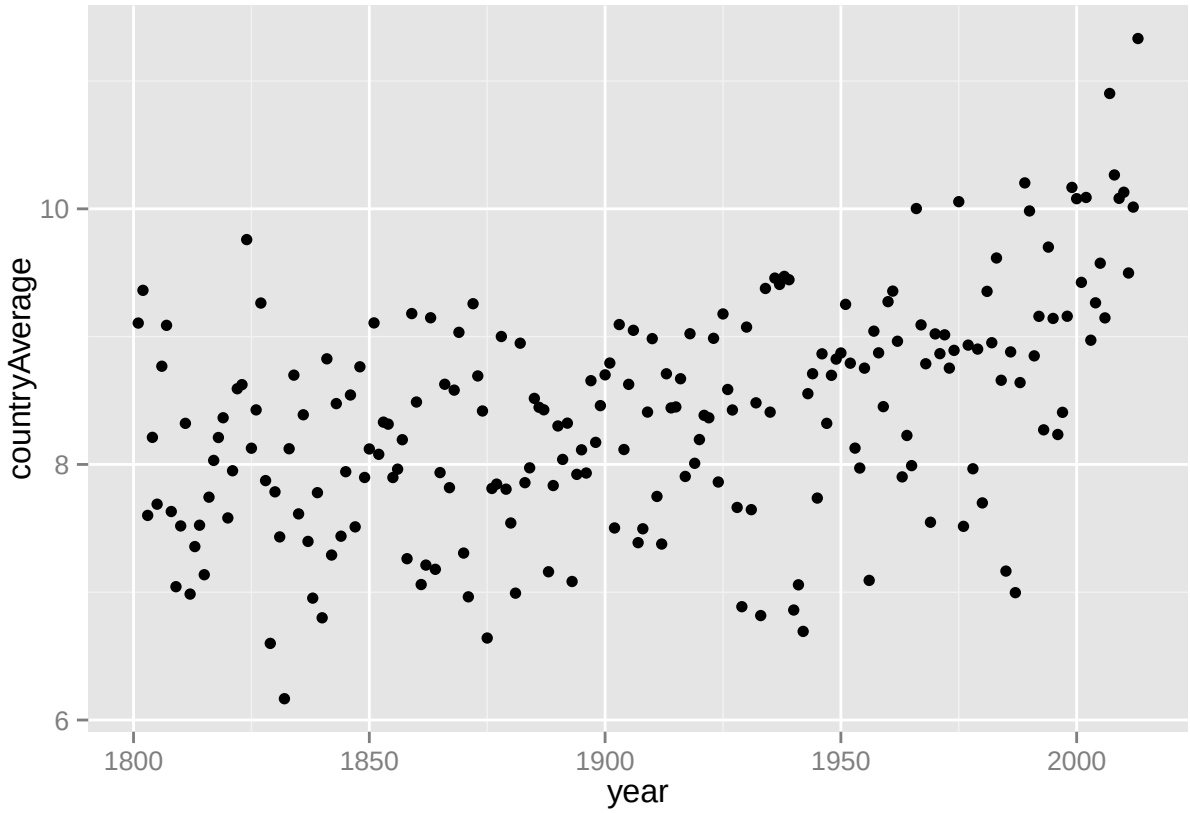
It seems that there is some problem with data availability before 1800. We shall filter out all the measurements before this year to see more clear picture. Let's use `filter` again

```
temperature_ukraine  <- temperature_ukraine %>%
                        filter(year > 1800)
```

... and try to visualise

```
yearly_ukraine_temp <- temperature_ukraine %>%
                        group_by(year) %>%
                        summarise(countryAverage = mean(AverageTemperatureCelsius))

ggplot(data = yearly_ukraine_temp, aes(x = year, y = countryAverage)) +
      geom_point()
```

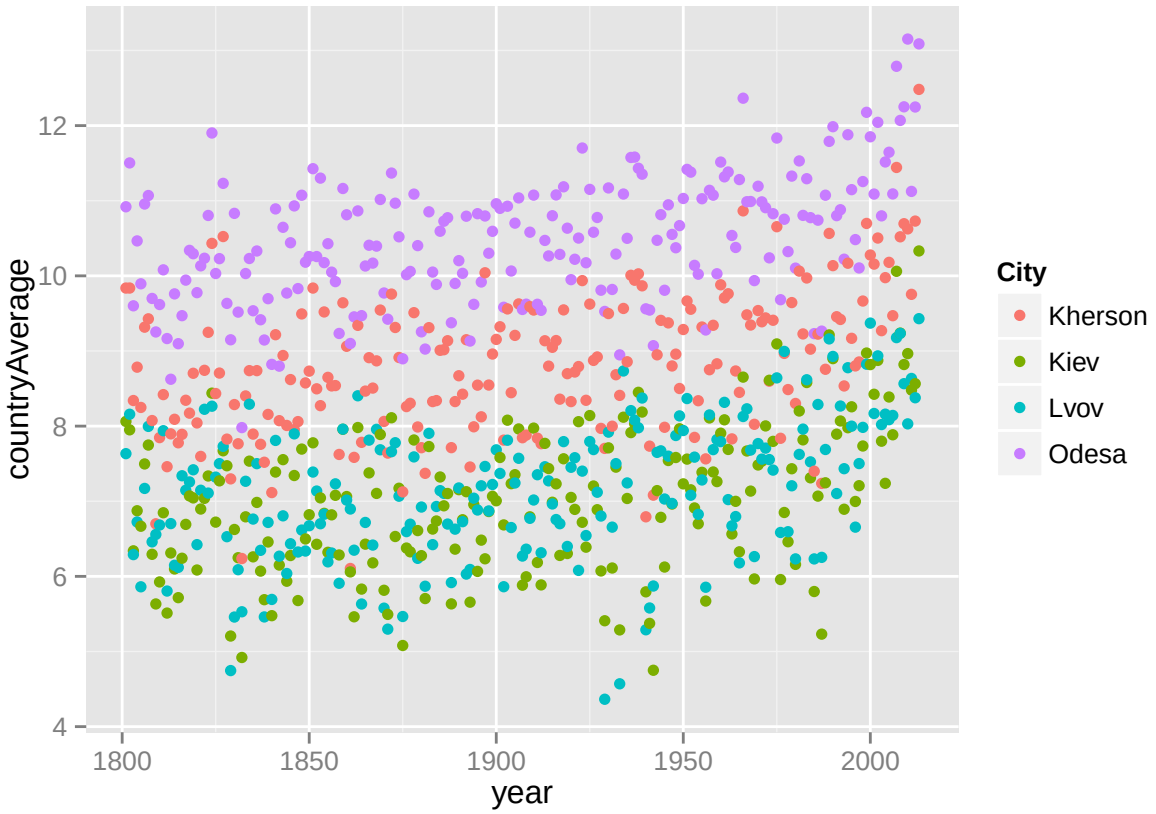


Let's see how much each city influences this trend

```
yearly_ukraine_temp <- temperature_ukraine %>%
                        group_by(year, City) %>%
                        summarise(countryAverage = mean(AverageTemperatureCelsius))
```

adding `City` as a colour component will add this info to our graph

```
ggplot(data = yearly_ukraine_temp, aes(x = year, y = countryAverage, group = City, colour = City)) +
      geom_point()
```



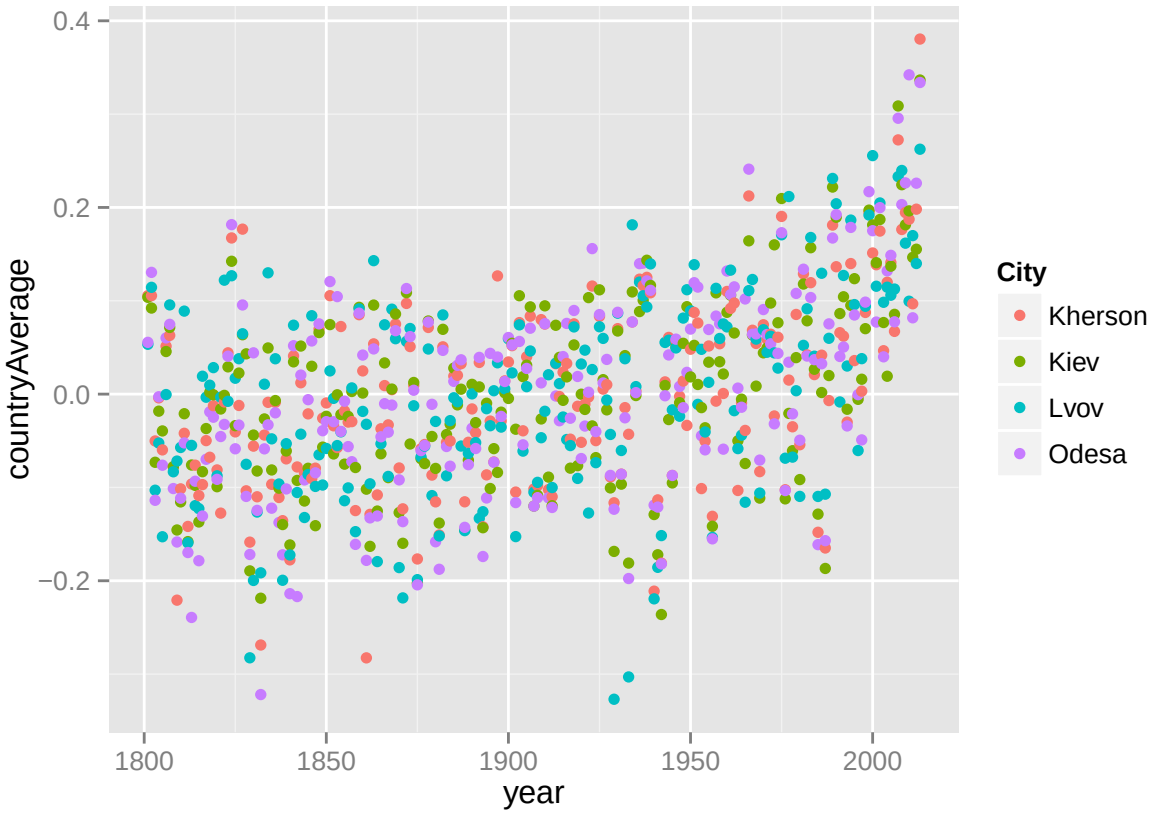
Let's scale all the distributions by average of each city and divide by the standard deviation of each city We shall use `mutate` to store scalled temperature, which is a result of subtracting average of each group and dividing them by the standard deviation

```
scaled_temperature_ukraine <- temperature_ukraine %>%
                              group_by(City) %>%
                              mutate(scaledTemperature = scale(AverageTemperatureCelsius)[,1])
```

Now let's plot the same

```
yearly_ukraine_temp <- scaled_temperature_ukraine %>%
                        group_by(year, City) %>%
                        summarise(countryAverage = mean(scaledTemperature))
```

```
ggplot(data = yearly_ukraine_temp, aes(x = year, y = countryAverage, group = City, colour = City)) +
      geom_point()
```

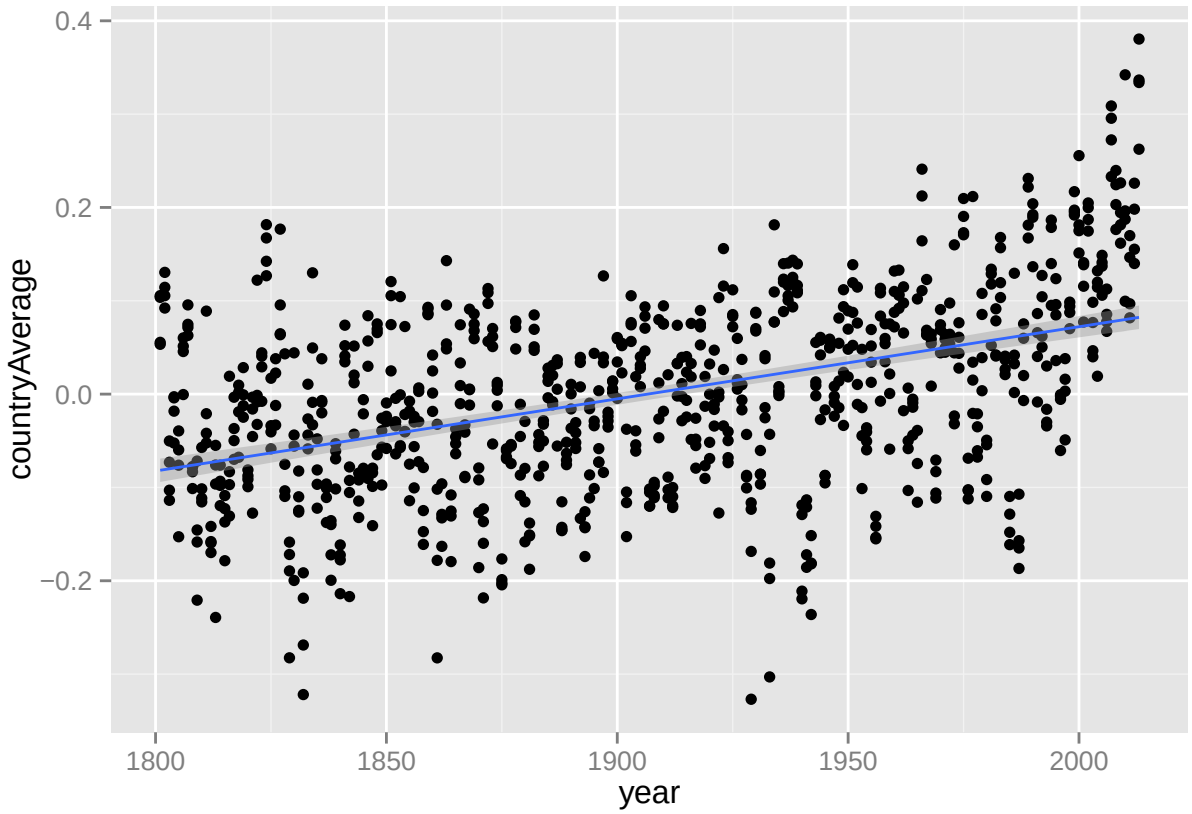


This looks much better, I mean much worse in terms of global warming of course, as it seems that the trend is indeed present.

Regression

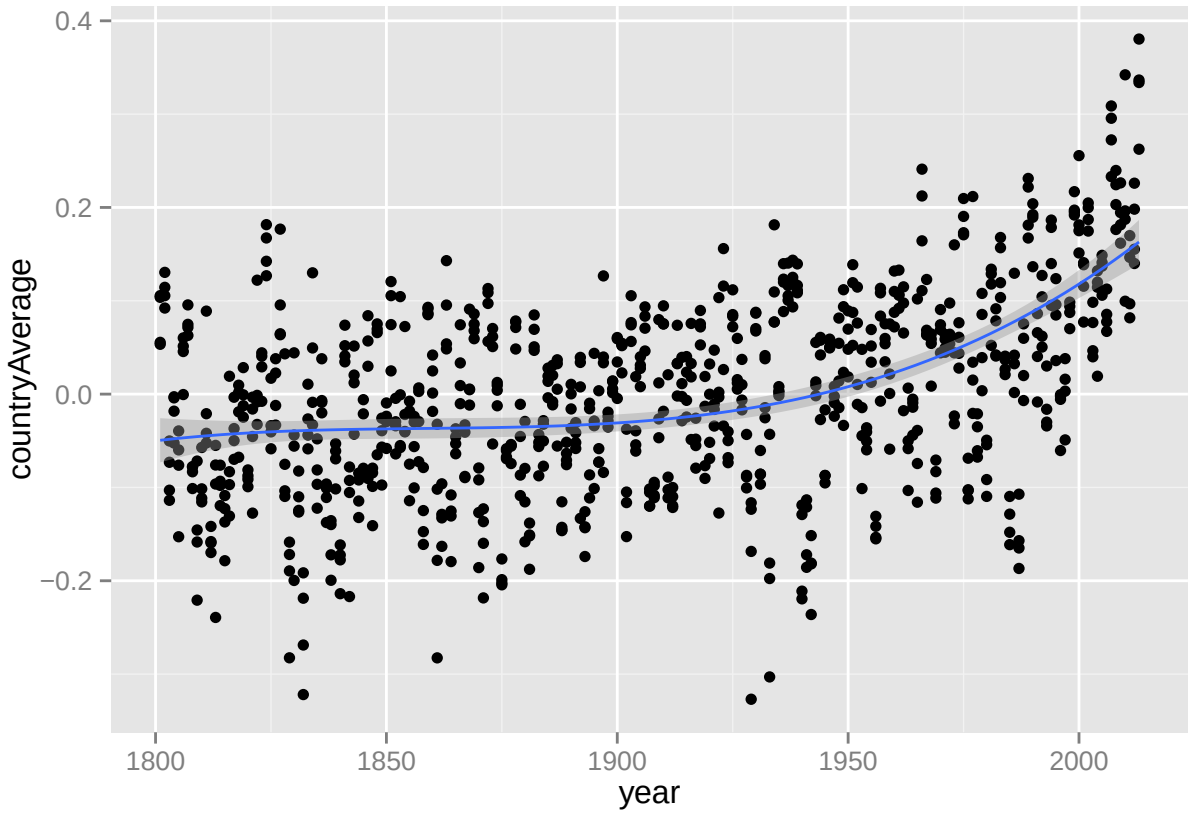
Let's add a regression line to the plot above. For that we will use `geom_smooth` function that fits regression to the data.

```
ggplot(data = yearly_ukraine_temp, aes(x = year,  y = countryAverage)) +
      geom_point() +
      geom_smooth(method = lm)
```



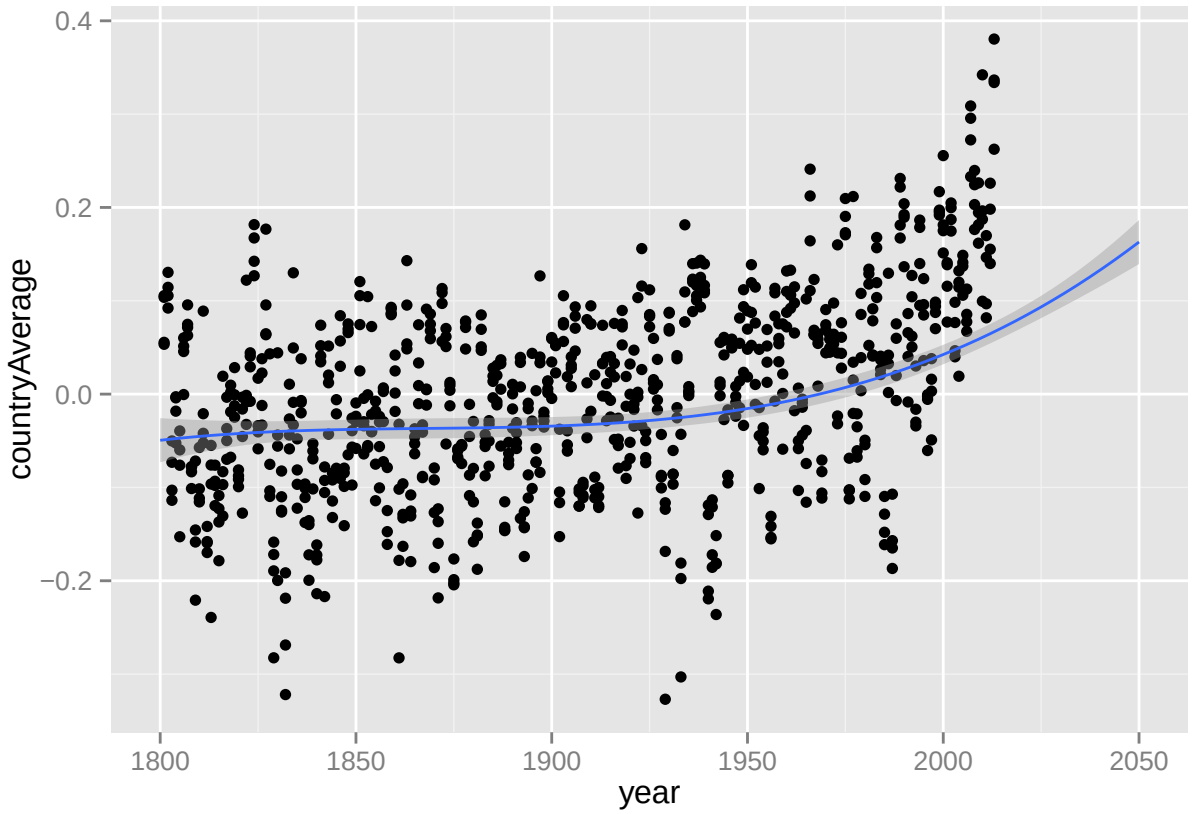
We might want to fit a polynomial trend instead of linear (it is too linear), this is how it is done:

```
ggplot(data = yearly_ukraine_temp, aes(x = year,  y = countryAverage)) +
      geom_point() +
      geom_smooth(method = lm, formula = y ~ splines::bs(x, 3))
```



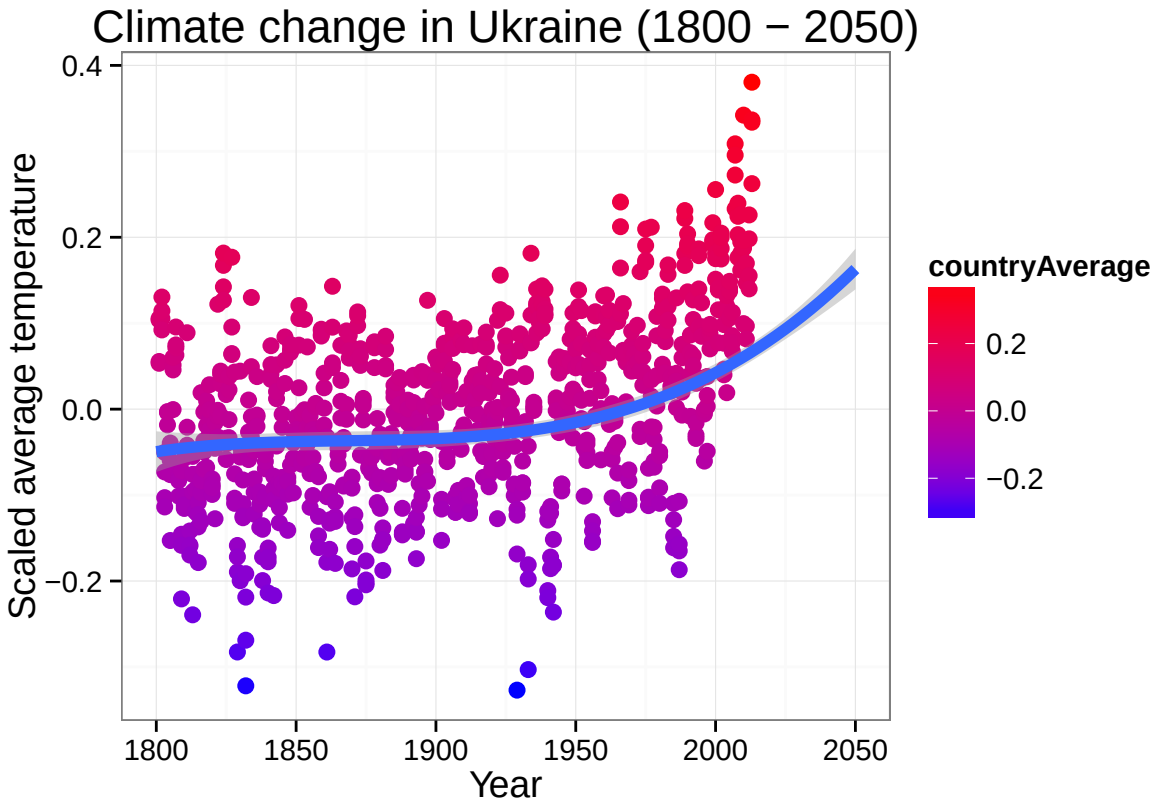
Also as we want to get some prediction we should fix the time range from 1800 to 2050 usign `xlim` function and set argument `fullrange` in `geom_smooth` to `TRUE`

```
ggplot(data = yearly_ukraine_temp, aes(x = year,  y = countryAverage)) +
      geom_point() +
      xlim(1800, 2050) +
      geom_smooth(method = lm, formula = y ~ splines::bs(x, 3), fullrange = TRUE)
```



Nice, we are almost there, quickly adding some bells and whistles for the graph to look more fancy

```
ggplot(data = yearly_ukraine_temp, aes(x = year,  y = countryAverage)) +
      geom_point(aes(colour = countryAverage), size = 3) +
      xlim(1800, 2050) +
      geom_smooth(size = 2, method = lm, formula = y ~ splines::bs(x, 3), fullrange = TRUE) +
      labs(title = 'Climate change in Ukraine (1800 - 2050)',
                        x = 'Year',
                        y = 'Scaled average temperature') +
      theme_bw(base_size = 14) +
      scale_colour_gradient(low = "blue", high = "red")
```



Now, what about interpretation of this figure? We can see that by 2050 our scaled averaged temperature will differ by at least 0.2 standard deviations. 1 standard deviation of yearly average temperature in Ukraine is about 7.5 degrees, thus, 0.2*7.5 is about 1.5 degrees warmer on average per year.

At this level, expected within 40 years, the hot European summer of 2003 will be the annual norm. Anything that could be called a heatwave thereafter will be of Saharan intensity. Even in average years, people will die of heat stress. Global warming, our future